Technical Disclosure Commons

Defensive Publications Series

November 09, 2017

EARLY SPAM DETECTION USING PARTIALLY AVAILABLE CONTENT ITEM

Vladimir Rychev

Bartlomiej Wolowiec

EARLY SPAM DETECTION USING PARTIALLY AVAILABLE CONTENT ITEM

A content item service may allow users to upload content items (e.g., videos, songs, audiobooks, images, documents, etc.) on the content item service. Such content items may be streamed or otherwise provided or rendered to various users. The content item service may perform various processing actions with regards to content items when users start uploading content items. For example, one of the processing actions may include transcoding a content item. Transcoding may involve converting a content item from one coded format (e.g., MP3, MP4, MPEG-2, MPEG-4, AVI, JPEG, etc.) to another coded format. A content item may be transcoded into multiple formats to allow users to use it on various types of devices and platforms, such as, on desktop computers, laptops, phones, televisions, gaming consoles, different web browsers and operating systems, etc. The content item service may transcode the content item into formats with differing qualities (e.g., high resolution, low resolution, etc.) for users to choose from. The content item service may also spend resources on other types of non-transcoding processing, such as, generating meaningful thumbnails, analyzing content items to study content data, comparing with copyrighted content, etc.

Some users may upload spam items on the content item service. For example, spam items may include content that are undesired, vastly repetitive, misleading, deceptive, etc. Upon completion of a content item upload, the content item service may perform content analysis and detect that the content item is a spam item. Subsequently, the content item service may take a restrictive action according to its spam handling policies, such as, make the content item unavailable for users, deactivate the content item, remove it from the content item service, etc. However, by the time the content item may be detected as spam, significant amount of resources and cost may have been spent on performing the various processing actions. The cost and efforts

involved with processing a spam item may be unnecessary, wasteful and strenuous given the content item may be ultimately restricted from use.

We propose a mechanism for early detection of a spam item by utilizing partially available content item in advance of spending significant resources on fully processing the content item. For example, once the user initiates a content item upload, a portion of the content item may be received by the content item server and be analyzed to determine whether the content item is a likely spam item. As soon as a specified portion of the content item is received, multiple features may be extracted from the received portion. Since the received portion is a fraction of the full content item, the analysis on the received portion may require much less computational resources than performing an analysis on the full content. Since it is possible to analyze the received portion early in the uploading process (e.g., immediately after initiation of upload, prior to upload complete, etc.), a determination may be made whether it is necessary to devote additional resources for performing full scale processing. If the partially available content item indicates that the content item may be a likely spam item, then the content item service may initially perform only limited processing of the content item and generate a single, cheaper version of the transcode. By doing so, wasteful processing costs may be avoided for a content item that is indeed a spam item and would be eventually restricted. Subsequently, full content analysis may be performed on the single transcode and a determination can be made as to whether the content item is in fact a spam item, based on which a further restrictive action may be taken.

Figure 1 depicts a flow diagram of a method for early detection of spam based on partially available content item. First, at step 101, a content item service may detect initiation of a content item upload by a user on the content item service. For example, a user may initiate

uploading a content item on a content item service by selecting a content item from a device of the user (e.g., a smart phone, a PDA, a laptop, a personal computer, etc.). The user may be logged into the content item service using an account associated with the user. The content item may be a video file, an audio file, a word document, an image file, etc. The content item may be large, medium, small, etc. Portions of the content item may reach a server of the content item service in increments. The user's device may send an indication (e.g., a signal, a message, etc.) to the content item service when the user initiates the upload. The content item service may detect the indication immediately, or prior to or at the same time as receiving initial portions of the content item. The content item service may detect the initiation of a content item upload by the user as soon as either the indication or an initial portion of the content item is received. The initiation may be detected prior to completion of the content item upload.

Next, at step 102, a portion of a content item may be received. For example, a user may initiate uploading a 500MB movie file as a content item. A fraction (e.g. 1MB, 2MB, etc.) of the 500MB file may be received by the content item service at a time. The content item service may receive each portion of the content item gradually. Upon detecting initiation of the content item, the content item service may wait to receive a portion of the content item. The portion may be specified as a predetermined threshold amount. The threshold may be a fixed byte, number of frames in a video, a percentage of the total size of the content item, or a combination. For example, the threshold may be 10 MB, or 10% of the file, or whichever is received first. The threshold may vary depending on other factors, for example and not limited to, the size of the content item being uploaded, available bandwidth, type of content item, etc. The content item server may detect when the specified portion matching the threshold is received. The content

item may further wait for a second or more portions of the content item to be received before proceeding to the next operations.

At step 103, the content item service may determine, based on features of the received portion of the content item, that the content item is a likely spam item. For example, upon detection of receiving the specified threshold portion of the content item, the content item server may analyze the received portion to identify the likelihood of the user uploading a spam item. The content item may be classified as a likely spam item or not a likely spam item based on the analysis of the partially available content item (e.g. the received portion). In one example, the content item may be classified as a likely spam item. The analysis of the partially available content item may take various factors into consideration when classifying the user's content item as a likely spam item. For example, the factors may include but are not limited to, partial metadata (e.g., resolution, frame rate, bitrate, codecs, etc.) of the available portion, a sample of individual decoded frames, partially decoded data (e.g., transformation vectors, although not the actual image frames) raw data stream, etc.

The analysis of the partially available content item may be performed by running a partial content classifier. The partial content classifier may specify a certain threshold or criteria for classifying the user's content item as a likely spam item. For example, if 99% of the received portion matches with identified spam materials, the partial content classifier may identify the current content item as a likely spam item. In one example, the partial content classifier may be run on a different portion of the content item than one that has already been analyzed if there is a serious concern regarding the already analyzed portion. In another example, the partial classifier may identify the current content item as a likely spam item if the first 10MB of the content item is identical to the next 10 MB, etc. In other examples, the partial content classifier may use a

machine learning algorithm (e.g., a neural network) that has been trained on a number of factors, including those identified in the preceding paragraph. When the machine learning algorithm is provided with the partially available content item, it may provide a determination whether the content item is a likely spam item or not. In one example, the partial content classifier may determine that the content item is a likely spam item based on the partially available content item.

Subsequently, at step 104, the content item may be converted into a single file. For example, once the content item is determined as a likely spam, the content item service may limit the processing of the content item. The limited processing may include converting (i.e., transcoding) the content item into a single file instead of transcoding the content item into multiple files with various formats. In one example, the single file may be a low resolution file. The limited processing may also include abandoning various other transcoding actions that may have already started prior to determining the content item as a likely spam item, deprioritizing or abandoning any other non-transcoding processing of the content item, such as, comparing with copyrighted materials, etc. The single file may be published on the content item service fast and made available for users. The single file may be a format that is usable by majority of devices and platforms used by various users.

At step 105, content analysis may be performed on the single file. For example, uploading of the content item may be considered completed when the content item is converted into the single file and published. The content of the single file may be analyzed in full at this stage, including all portions of the content item. For example, the content item service may analyze the full content of a video file to determine whether the file includes vastly repeated portions throughout the file, or portions matching with identified spam materials, etc.

At step 106, the content item service may determine, based on the content analysis, that the content item is an actual spam item. For example, based on the results of the content analysis of the single file, the content item may at this point be classified as an actual spam item rather than a likely spam item.

Subsequently, at step 107, availability of the content item may be restricted on the content item service. For example, the content item service may deactivate the published content item, make the content item unavailable to regular users, only keep the content available for third parties, remove the content item from the content item service, etc. In an example, the content item may be kept on the content item service for a specified length of time. After the specified time, the content item may be removed. The content item service may allow the user uploading the content item an option to dispute that the content item is a spam. Further actions on the content item may be decided based on spam handling policies of the content item service.

In an implementation, if the content item is identified as a not likely spam based on partially available content tem, the content item processing may continue as is. That is, the content item may be transcoded into various formats, and any additional non-transcoding processing may continue to be performed. Content analysis may be performed on all of the transcoded formats of the content item or any one of the transcoded formats (e.g. on a transcode that meets a minimum resolution, etc.). Appropriate (e.g., restrictive, non-restrictive) actions may be taken based on the result of content analysis (e.g., whether actual spam or not actual spam).

In another implementation, after the content item is determined as a likely spam item and has been converted into a single file, if it is determined based on the content analysis that the content item is not an actual spam item, further processing steps may be taken. The further processing steps may include transcoding the content item into various desired formats, and

starting or resuming other non-transcoding processing actions, such as data analysis, copyright validation, etc. The content item may be fully processed and available to various users on widespread formats. In an example, no immediate restrictive actions may be necessary if it is determined that the content item is not a spam item.

The mechanism described herein allows early detection of spam by utilizing partially available content item in advance of spending significant resources on fully processing the content item. Because the mechanism allows limited processing of content item that is marked as a likely spam item, valuable resources are not spent on full processing of the content item and the content item is not transcoded into various formats when there is likelihood that the content item may ultimately be restricted or removed from the content item service. Additionally, since the content item still goes through some minimal processing and is transcoded into at least a single file, this mechanism does not introduce latency issues. The content item can be also deleted sooner in the process because the partially available portion of the content item can indicate the item is likely spam sooner in the process.

ABSTRACT

A mechanism for early detection of a spam item by utilizing partially available content item in advance of spending significant resources on fully processing the content item is disclosed. Upon detection of initiation of a content item uploaded by a user on a content item service, a portion of the content item may be received. Based on features of the received portion of the content item, it may be determined that the content item is a likely spam item. Subsequently, limited processing maybe performed by converting the content item may into a single file and content analysis may be performed on the single file. Based on the content analysis, it may be determined that the content item is an actual spam item and the availability of the content item may be restricted on the content item service.

Keywords: video, content, partial content item, transcode, content analysis, spam.

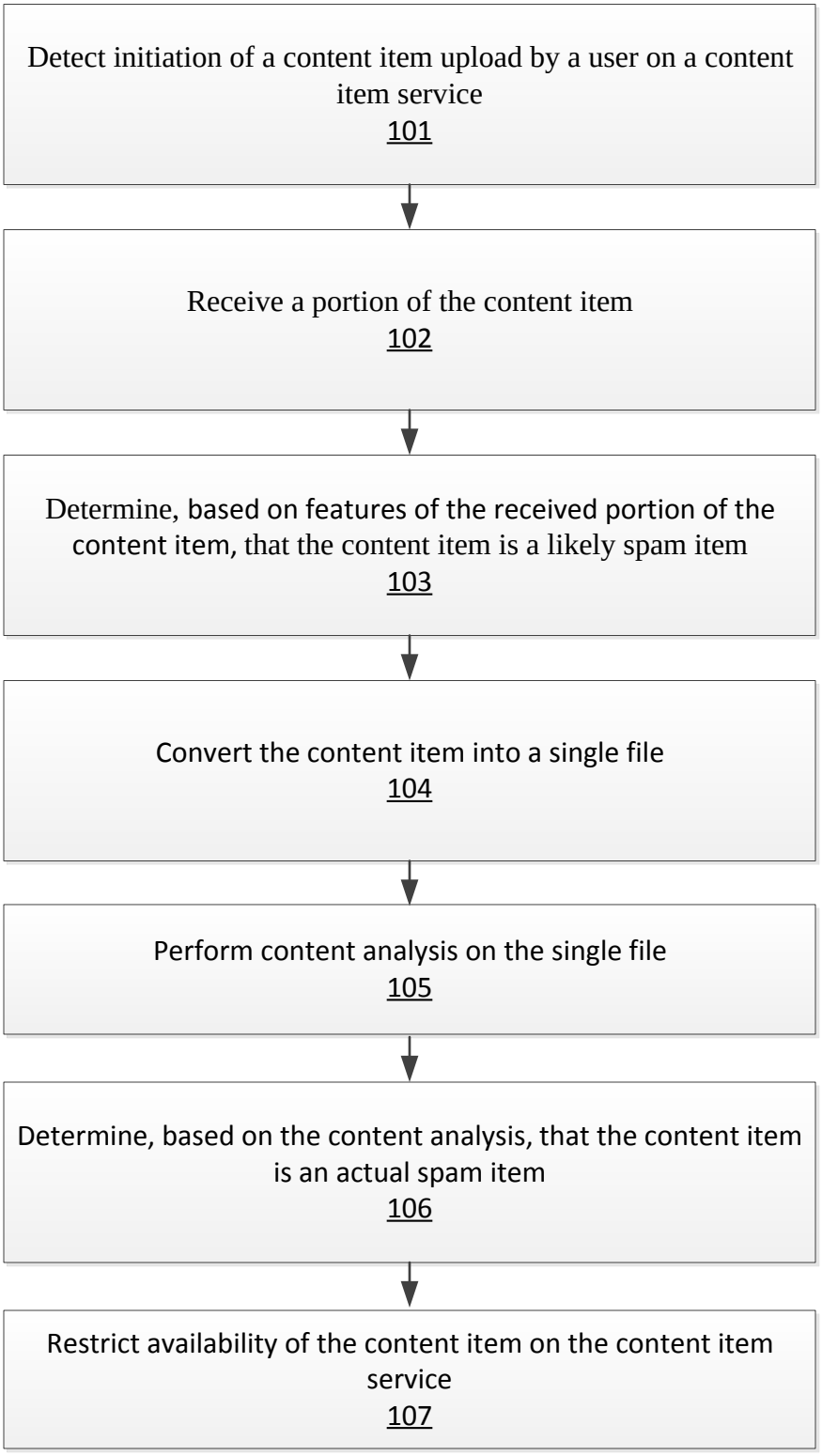
Detect initiation of a content item upload by a user on a content item service
101
Receive a portion of the content item
102
Determine, based on features of the received portion of the content item, that the content item is a likely spam item
103
Convert the content item into a single file
104
Perform content analysis on the single file
105
Determine, based on the content analysis, that the content item is an actual spam item
106
Restrict availability of the content item on the content item service
107

FIG. 1